

Low-rank Interaction Contingency Tables

Geneviève Robin¹

Julie Josse¹, Éric Moulines¹, and Sylvain Sardy²

1. Center of Applied Mathematics, École Polytechnique, France

2. Department of Mathematics, Université de Genève, Switzerland

March 8, 2017

Abstract

Log-linear models are popular tools to analyze contingency tables, particularly to model row and column effects as well as row-column interactions in two-way tables. In this paper, we introduce a regularized log-linear model designed for denoising and visualizing count data, which can incorporate side information such as row and column features. The estimation is performed through a convex optimization problem where we minimize a negative Poisson log-likelihood penalized by the nuclear norm of the interaction matrix. We derive an upper bound on the Frobenius estimation error, which improves previous rates for Poisson matrix recovery, and an algorithm based on the alternating direction method of multipliers to compute our estimator. To propose a complete methodology to users, we also address automatic selection of the regularization parameter. A Monte Carlo simulation reveals that our estimator is particularly well suited to estimate the rank of the interaction in low signal to noise ratio regimes. We illustrate with two data analyses that the results can be easily interpreted through biplot visualization. The method is available as an R code.

Keywords: count data, low-rank matrix recovery, dimensionality reduction, EM algorithm, quantile universal threshold

1 Introduction

High dimensional count data are collected in many scientific and engineering tasks including image processing (Luisier et al., 2011), single-cell RNA sequencing (Pierson and Christopher, 2015; O. Stegle and Marioni, 2015) and ecological studies (Peres-Neto et al., 2016). In this latter field, scientists often aim at understanding how species interact with different sites or biological environments. The data consist of contingency tables collecting the abundance of species across sampling sites, along with explanatory covariates providing additional information about species and environmental traits. The goal is then to uncover the interactions that cause some species to thrive or decay in particular environments. Consider an $m_1 \times m_2$ observation matrix of counts $Y = (y_{ij})$ with independent cells of means $\mathbb{E}[y_{ij}] = \exp(\bar{x}_{ij})$. Log-linear models (Agresti, 2013; Christensen, 2010) aim at describing the structure of the mean matrix $\bar{X} = (\bar{x}_{ij})$. The saturated model can be written as follows:

$$\bar{x}_{ij} = \bar{\alpha}_i + \bar{\beta}_j + \bar{\Theta}_{ij}, \quad (1.1)$$

where $\bar{\alpha}_i$ (resp. $\bar{\beta}_j$) accounts for the main additive effect of row i (resp. column j) while $\bar{\Theta}_{ij}$ is a row-column interaction term. In the ecological application mentioned above, parameters $\bar{\alpha}_i$ correspond to environment effects, while the $\bar{\beta}_j$ correspond to species effects. Model (1.1) is overparametrized but can be restricted using the log-bilinear model, also known as the RC(K) model (RC for row-column) or the generalized additive main effects and multiplicative interaction (GAMMI) model (Goodman, 1985; de Falguerolles, 1998; Gower et al., 2011; Fithian and Josse, 2017). The GAMMI model assumes that the interaction matrix $\bar{\Theta}$ has fixed rank $K \leq \min(m_1 - 1, m_2 - 1)$. This procedure requires to estimate the rank of the interaction from data, which is a difficult task. The estimation of the means $\exp(\bar{x}_{ij})$ is then often done by minimizing a negative Poisson log-likelihood,

defined for $X \in \mathbb{R}^{m_1 \times m_2}$ by

$$\Phi_Y(X) = -\frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} (y_{ij} x_{ij} - \exp(x_{ij})). \quad (1.2)$$

In this paper, we propose a regularized generalized bilinear model, named GAMMIT that stands for *generalized additive main effects and multiplicative interaction thresholded*, for denoising and visualization of two-way contingency tables. It extends the log-bilinear model (1.1) in two ways. First, it allows to account for general covariates instead of only row and column effects. Secondly, it improves on the maximum likelihood estimation by including a regularization through the nuclear norm penalty of the interaction matrix. Let us explain the former point in details to show that it allows subtle incorporation of all the variability explained by the known covariates. Let $R \in \mathbb{R}^{m_1 \times K_1}$ (resp. $C \in \mathbb{R}^{K_2 \times m_2}$) be matrices of known row (resp. column) covariates, and $\bar{\alpha} \in \mathbb{R}^{K_1 \times m_2}$ (resp. $\bar{\beta} \in \mathbb{R}^{m_1 \times K_2}$) be unknown parameters. We model the matrix \bar{X} as follows:

$$\bar{X} = R\bar{\alpha} + \bar{\beta}C + \bar{\Theta}. \quad (1.3)$$

In our example in ecology, the row features R embed geographical information about the environments (slope, temperature, etc.), and C codes physical traits about species (height, mass, etc.). If $\bar{\alpha}$ has constant rows, $\bar{\alpha}_{ij}$ denotes the effect of the j -th environment covariate for all i . Then $\bar{\alpha}_{ij}$ will be large if large values of the j -th environment covariate, say the temperature, lead to large abundances of all species. Similarly if $\bar{\beta}$ has constant columns, $\bar{\beta}_{ij}$ denotes the effect of the i -th species covariate for all j . Then $\bar{\beta}_{ij}$ will be large if large values of the i -th species covariate, say the weight, lead to large abundances in all environments. The matrices $\bar{\alpha}$ and $\bar{\beta}$ can also incorporate interactions, and will not be constant in this case. The coefficient $\bar{\alpha}_{ij}$ will then be large if large values of the j -th environment covariate leads to large abundances of species i . Similarly the coefficient $\bar{\beta}_{ij}$ will be large if large

values of the i -th species covariate leads to large abundances in environment j . We consider this latter general case in this article. The matrix $R\bar{\alpha}$ therefore models the main effects of row covariates plus the interaction between columns and row covariates. The matrix $\bar{\beta}C$ models the main effects of the columns covariates plus the interaction between columns covariates and rows. Then $\bar{\Theta}$, that we want to uncover, corresponds to the interaction between species and environments, which is unexplained by the known covariates R and C , but nonetheless influences the observations.

The paper is organized as follows. In Section 2, we define our estimator through the minimization of a negative Poisson log-likelihood term (1.2) penalized by the nuclear norm of the interaction matrix $\bar{\Theta}$. Then, we propose an algorithm based on the alternating descent method of multipliers (ADMM) to solve the convex optimization problem and compute this estimator. Under mild assumptions on the true parameter matrix \bar{X} , we derive in Section 3 an upper bound for the Frobenius estimation error of this estimator that holds with high probability. Another main contribution is to propose in Section 4 two methods to choose the regularization parameter. The first one is based on cross-validation and thus requires to define an EM algorithm to deal with missing values; as an aside, it gives a new method to impute contingency tables. The second approach is inspired by the work of Giacobino et al. (2016) on quantile universal thresholds (QUT). QUT also interestingly yields a new test of independence that can be used as an alternative to the χ^2 test. Finally, we illustrate the performance of our procedure in Section 5 on synthetic data, and the analysis of an ecological data set is detailed in Section 6. We highlight the interpretability of the results using biplots visualization. The experiments presented in this article are reproducible and available as an R code (R Core Team, 2016) on Github at <https://github.com/genevievevelrobin/GAMMIT>.

Related approaches for Poisson matrix recovery and dimensionality reduction can be

embedded within the framework of low-rank exponential family estimation (Collins et al., 2001; de Leeuw, 2006; Li and Tao, 2013; Josse and Wager, 2016; Liu et al., 2016). Existing models impose low-rank either to the natural parameter matrix with cells \bar{x}_{ij} (Collins et al., 2001) or to the mean with cells $\exp(\bar{x}_{ij})$ (Liu et al., 2016; Josse and Wager, 2016). Some procedures maximize a Poisson log-likelihood subject to a nuclear norm penalty, leading to non-quadratic and non-separable problems. Proposed optimization approaches include iterative partial updates of the parameters (Salmon et al., 2014) and augmented Lagrangian methods (Figueiredo and Bioucas-Dias, 2010; Chambolle and Pock, 2011; Jeong et al., 2013). The theoretical performances of nuclear norm penalized estimators have been studied in Cao and Xie (2016), where the authors prove uniform bounds on the empirical error risk by extending results of compressed sensing and 1-bit matrix completion (Raginsky et al., 2010; Davenport et al., 2012). Our results improve on the rates reported in these works. Poisson matrix denoising has also been considered through singular value shrinkage, extending the Gaussian setting (Shabalin and Nobel, 2013; Gavish and Donoho, 2014b,a; Josse and Sardy, 2015a). Bigot et al. (2016) studied optimal singular value shrinkage for low-rank matrix denoising in the exponential family, while Liu et al. (2016) suggested a new shrinker for covariance matrix estimation. None of the methods reviewed so far accounts for effects of known covariates. Attempts to include row and column effects in matrix recovery and completion have nonetheless been made in the context of the Netflix challenge. Some are reviewed in Feuerverger et al. (2012), and Hastie et al. (2014) briefly addressed this issue through centering and scaling steps. The GAMMI model introduced by Goodman (1985) takes into account main additive effects and multiplicative interactions.

2 Method

The main idea is that we want to decompose the parameter matrix in $\bar{X} = \bar{X}_0 + \bar{\Theta}$, where \bar{X}_0 is explained by the known covariates while $\bar{\Theta}$ is not, with $\bar{X}_0 \perp \bar{\Theta}$ in the sense of the trace scalar product. To do so, we first need to introduce some notations. Let \mathcal{V}_1 (resp. \mathcal{V}_2) be the linear span of the columns of R (resp. rows of C) in model (1.3) of dimension K_1 (resp. K_2). Define $\Pi_1 \in \mathbb{R}^{m_2 \times m_2}$ (resp. $\Pi_2 \in \mathbb{R}^{m_1 \times m_1}$) the orthogonal projection matrices on subspace \mathcal{V}_1 (resp. \mathcal{V}_2). For ease of notation we call I the identity matrices of $\mathbb{R}^{m_1 \times m_1}$ and $\mathbb{R}^{m_2 \times m_2}$. Let $\bar{X}_{.,j}$ be the j -th column of \bar{X} , we write $\bar{X}_{.,j} = \Pi_2 \bar{X}_{.,j} + (I - \Pi_2) \bar{X}_{.,j}$. Similarly $\bar{X}_{i,.} = \bar{X}_{i,.} \Pi_1 + \bar{X}_{i,.} (I - \Pi_1)$. The parameter matrix \bar{X} can now be decomposed in $\bar{X} = \bar{X}_0 + \bar{\Theta}$ with $\bar{\Theta} = (I - \Pi_2) \bar{X} (I - \Pi_1)$. We now identify

$$\begin{aligned} \bar{X}_0 &= \bar{X} \Pi_1 + \Pi_2 \bar{X} - \Pi_2 \bar{X} \Pi_1 \\ \bar{X}_0 &= R\bar{\alpha} + \bar{\beta}C \end{aligned} \tag{2.1}$$

As detailed in the introduction, the interpretation of X_0 is subtle since it contains the effects of the covariates, namely $R\alpha$ the main effects of the row covariates plus their interaction with the columns and βC the main effects of the column covariates plus their interaction with the rows. Note that $\Pi_2 \bar{X} \Pi_1$ denotes the interaction between row covariates and column covariates and has to be subtracted from \bar{X}_0 since it is contained in both $\bar{X} \Pi_1$ and $\Pi_2 \bar{X}$. Let \mathcal{V} be the linear span of dimension $K_1 m_2 + m_1 K_2 - K_1 K_2$ of matrices of the form $\Pi_2 \bar{X} + \bar{X} \Pi_1 - \Pi_2 \bar{X} \Pi_1$. We define the orthogonal projection operator \mathcal{T} on \mathcal{V}^\perp . In other words $\mathcal{T} : \bar{X} \mapsto (I - \Pi_2) \bar{X} (I - \Pi_1)$, which implies $\bar{\Theta} = \mathcal{T}(\bar{X})$ and $\bar{X}_0 \perp \bar{\Theta}$. In the sequel, for $X \in \mathbb{R}^{m_1 \times m_2}$ we write the Schatten s -norm (Bhatia and Kittaneh, 2000) with $s \in [1, \infty)$ $\|X\|_{\sigma,s} = (\sum_{i=1}^{m_1 \wedge m_2} \sigma_i^s(X))^{1/s}$, and $\|X\|_{\sigma,\infty}$ the largest singular value of X . We finally define our estimator for a given regularization parameter λ as the minimizer of the

penalized negative log-likelihood:

$$\hat{X}_\lambda = \underset{X}{\operatorname{argmin}} \quad \Phi_Y^\lambda(X), \quad (2.2)$$

$$\Phi_Y^\lambda(X) = \Phi_Y(X) + \lambda \|\mathcal{T}(X)\|_{\sigma,1}. \quad (2.3)$$

We now describe an optimization algorithm to solve (2.2). Consider the following assumption.

- H 1.** (i) *There exist $\underline{\mu} > 0$ and $\bar{\mu} < \infty$ such that for all $i, j \in [m_1] \times [m_2]$, $\underline{\mu} \leq \mathbb{E}[y_{ij}] \leq \bar{\mu}$.*
(ii) *There exist $\underline{\sigma} > 0$ and $\bar{\sigma} < \infty$ such that for all $i, j \in [m_1] \times [m_2]$, $\underline{\sigma}^2 \leq \operatorname{Var}(y_{ij}) \leq \bar{\sigma}^2$.*

Under **H 1** and setting $\underline{\gamma} = \log(\underline{\mu})$ and $\bar{\gamma} = \log(\bar{\mu})$, the entries of \bar{X} satisfy $\underline{\gamma} \leq \bar{x}_{ij} \leq \bar{\gamma}$ for all $i, j \in \{1, \dots, m_1\} \times \{1, \dots, m_2\}$. In the sequel we write $\mathcal{K} = [\underline{\gamma}, \bar{\gamma}]^{m_1 \times m_2}$. The parameter set \mathcal{K} is compact and Φ_Y^λ is strongly convex on \mathcal{K} , which guarantee existence and uniqueness of the solution of (2.2). Assumption **H 1** is common in the Poisson matrix denoising and completion literature. We solve (2.2) by using the alternating directions method of multipliers (ADMM, Glowinski and Marrocco (1974)), whose convergence stems from Boyd et al. (2011, Theorem 3.2.1). We solve the following reparametrized program:

$$\underset{X \in \mathcal{K}, \Theta \in \mathcal{K}_\mathcal{T}}{\operatorname{argmin}} \quad \Phi_Y(X) + \lambda \|\Theta\|_{\sigma,1} \quad \text{s.t.} \quad \mathcal{T}(X) - \Theta = 0, \quad (2.4)$$

where $\mathcal{K}_\mathcal{T}$ is the image of set \mathcal{K} by the projector \mathcal{T} , and is therefore also compact. The reparametrized problem (2.4) is strongly convex on a compact set, linearly constrained and separable in X and Θ . ADMM is a variant of the augmented Lagrangian method of multipliers which solves the dual problem through iterated partial updates. The augmented Lagrangian indexed by τ is written

$$\mathcal{L}_\tau(X, \Theta, \Gamma) = \Phi_Y(X) + \lambda \|\Theta\|_{\sigma,1} + \langle \Gamma, \mathcal{T}(X) - \Theta \rangle + \frac{\tau}{2} \|\mathcal{T}(X) - \Theta\|_{\sigma,2}^2, \quad (2.5)$$

where $\langle ., . \rangle$ denotes the trace scalar product on $\mathbb{R}^{m_1 \times m_2}$. ADMM consists in separate updates of the primal variables X , Θ and of the dual variable Γ to maximize (2.5) according to the following rules

$$\begin{aligned} X^{k+1} &= \operatorname{argmin}_{X \in \mathcal{K}} \mathcal{L}_\tau(X, \Theta^k, \Gamma^k) \\ \Theta^{k+1} &= \operatorname{argmin}_{\Theta \in \mathcal{K}_\tau} \mathcal{L}_\tau(X^{k+1}, \Theta, \Gamma^k) \\ \Gamma^{k+1} &= \Gamma^k + \tau(\mathcal{T}(X^{k+1}) - \Theta^{k+1}). \end{aligned} \tag{2.6}$$

The function Φ_Y and $\|\cdot\|$ are closed, proper and convex on $\mathbb{R}^{m_1 \times m_2} \rightarrow \mathbb{R}$. This guarantees the solvability of the minimization problems defined in the ADMM update (2.6). Moreover Φ_Y is differentiable, therefore the optimization in X can be done using gradient descent. The update of U can itself be done in closed form and involves singular value decomposition and thresholding (Cai et al., 2010):

$$\Theta^{k+1} = \mathcal{D}_{\lambda/\tau}(\mathcal{T}(X^{k+1}) + \Gamma/\tau),$$

where $\mathcal{D}_{\lambda/\tau}$ is the soft-thresholding operator of singular values at level λ/τ . To speed up ADMM to convergence, we implemented a warm-start strategy (Friedman et al., 2007; Hastie et al., 2015): we start by running the algorithm with $\lambda = \lambda_0(Y)$, the smallest value of the regularization parameter that sets the interaction to 0 (see Section 4); we then solve the optimization problem for decreasing values of λ , each time using the previous estimator as an initial value.

3 Statistical guarantees

In this section we present statistical guarantees on the Frobenius estimation error of estimator (2.2) under mild assumptions on the true parameter matrix \bar{X} . Our first result gives

an upper bound on the Frobenius estimation error of \hat{X}_λ that depends on the regularization parameter λ . Our second result gives a theoretical value of λ for which we control the estimation error with high probability. We denote by $\text{rk}(X)$ the rank of X , and $m = m_1 \wedge m_2$, $M = m_1 \vee m_2$, $d = m_1 + m_2$.

Theorem 3.1. *Assume **H 1** and $\lambda \geq 2 \|\nabla \Phi_Y(\bar{X})\|_{\sigma, \infty}$. Then*

$$\frac{\|\bar{X} - \hat{X}_\lambda\|_{\sigma, 2}^2}{m_1 m_2} \leq \lambda^2 / \underline{\mu}^2 m_1 m_2 (18 \text{rk}(\mathcal{T}(\bar{X})) + K_1 + K_2). \quad (3.1)$$

Proof. See Appendix A. □

We now discuss conditions under which $\lambda \geq 2 \|\nabla \Phi_Y(\bar{X})\|_{\sigma, \infty}$ holds with high probability, based on concentration inequalities for the largest eigenvalue of subexponential random matrices.

H 2. *There exists $\delta > 0$ such that for all $i, j \in [m_1] \times [m_2]$, $\mathbb{E}[\exp(|y_{ij}|/\delta)] < +\infty$.*

Theorem 3.2. *Under **H 1** and **H 2**, set $\lambda = 2c_\delta \bar{\sigma} \sqrt{2M \log(m_1 + m_2)/(m_1 m_2)}$ and assume $m_1 + m_2 \geq \max\{\delta^2(2\bar{\sigma}^2 \underline{\sigma}^2)^{-1}, (4\delta^2/\bar{\sigma}^2)^4\}$. Then with probability at least $1 - (m_1 + m_2)^{-1}$,*

$$\frac{\|\bar{X} - \hat{X}_\lambda\|_{\sigma, 2}^2}{m_1 m_2} \leq 4c_\delta \bar{\sigma}^2 / \underline{\mu}^2 \frac{M (18 \text{rk} \mathcal{T}(\bar{X}) + K_1 + K_2) \log(m_1 + m_2)}{m_1 m_2}, \quad (3.2)$$

where c_δ is a numerical constant that depends only on δ .

Proof. See Appendix B. □

4 Choice of λ

We propose two approaches for the automatic selection of λ : cross validation and a rule inspired by the universal threshold of [Donoho and Johnstone \(1994\)](#). We discuss their respective characteristics on practical cases in Section 5.

4.1 Cross-validation

The procedure consists in erasing a fraction of the observed values in Y , estimating a complete parameter matrix \hat{X}_λ for a range of λ values, and choosing the parameter λ that minimizes the error on the prediction of the removed values. This requires an estimation procedure that handles missing data. We derive an expectation-maximization (EM, [Dempster et al. \(1977\)](#)) algorithm to do so. We denote by Y_{obs} (resp. Y_{mis}) the observe (resp. missing, i.e. removed) entries of Y . The k th iteration of the EM algorithm goes as follows. In the E-step, we compute the expectation of the complete likelihood $\Phi_{(Y_{\text{obs}}, Y_{\text{mis}})}(X)$ with respect to the conditional distribution of the missing values Y_{mis} given the observed values Y_{obs} and the current parameter \hat{X}_λ^k . This boils down to replacing the missing entries Y_{mis} by their expected values. We obtain $Y_{\text{obs}}^{k+1} = Y_{\text{obs}}$ (unchanged) and $Y_{\text{mis}}^{k+1} = \exp(\hat{X}_{\lambda, \text{mis}}^k)$, where $\exp(X)$ denotes the cell-wise exponential of matrix X . In the M-step, we maximize the objective function $\Phi_{(Y^{k+1})}^\lambda(X)$ with respect to parameter X , giving $\hat{X}_\lambda^{k+1} = \arg \max_X \Phi_{(Y^{k+1})}^\lambda(X)$. This maximization step can be done using the ADMM algorithm described in [Section 2](#). These two steps are iterated until convergence. Repeating this procedure, say N times, for a grid of λ , we select the value of λ that minimizes the prediction square error $\text{PSE}_\lambda = 1/N \sum_{i=1}^N \left\| Y_{\text{mis}} - \hat{X}_{\lambda, \text{mis}}^{(i)} \right\|_2^2$, which is a proxy for the regularization parameter that minimizes $\left\| \bar{X} - \hat{X}_\lambda \right\|_{\sigma, 2}^2$. Notice that in the process, we have defined an algorithm to estimate \bar{X} from incomplete observations. This method can therefore also be seen as a matrix completion or single imputation method ([Little and Rubin, 2002](#)), and could be used as an alternative to existing techniques to complete contingency tables with missing values. This point definitely deserves further investigation.

4.2 Quantile Universal Threshold

Cross-validation is computationally intensive. It is well-suited to find a value of the regularization parameter λ with good prediction errors, but is not designed to estimate the rank of $\mathcal{T}(\bar{X})$. We suggest an alternative method to select λ , which is based on the work of [Giacobino et al. \(2016\)](#). They propose a generic approach to select regularization parameters for thresholding estimators based on the concept of a zero-thresholding statistic. Estimator (2.2) is a thresholding estimator in the sense that there exists a value $\lambda_0(Y)$ that depends on the data, for which the estimated interaction matrix is null, and the same estimate $\mathcal{T}(\hat{X}_\lambda) = 0$ is obtained for any $\lambda \geq \lambda_0(Y)$. $\lambda_0(Y)$ defines the zero-thresholding statistic for (2.2), that we derive in Proposition 1.

Proposition 1. *The interaction estimator $\mathcal{T}(\hat{X}_\lambda)$ associated to regularization parameter λ is null if and only if $\lambda \geq \lambda_0(Y)$, where $\lambda_0(Y)$ is the zero-thresholding statistic given by*

$$\lambda_0(Y) = (m_1 m_2)^{-1} \left\| \mathcal{T}(Y - \exp(\hat{X}_0)) \right\|_{\sigma, \infty},$$

where $\hat{X}_0 = \underset{X \in \mathcal{K}, \mathcal{T}(X)=0}{\operatorname{argmin}} \Phi_Y(X)$.

Proof. See Appendix C □

Note that in the log-bilinear model (1.1), \hat{X}_0 has an explicitly expression ([Kateri, 2014](#), Section 4.2). We propose to use the value of the zero-thresholding statistic $\lambda_0(Y)$ as the regularization parameter in our method, mainly because it has the nice property of setting the estimated interaction to 0 when the true interaction is indeed null. We now describe how it can be employed to test the null hypothesis $\mathbf{H}_0 : \mathcal{T}(\bar{X}) = 0$ against the alternative $\mathbf{H}_1 : \mathcal{T}(\bar{X}) \neq 0$, which boils down to testing if the measured covariates are sufficient to explain the observations. Suppose that for any λ we have access to the distribution

function $\mathbb{P}_{\mathbf{H}_0}(\lambda_0(Y) < \lambda)$. Proposition 1 ensures that model (2.2) gives $\mathcal{T}(\hat{X}_\lambda) = 0$ if and only if $\lambda \geq \lambda_0(Y)$. For any threshold λ , under \mathbf{H}_0 , the estimate $\mathcal{T}(\hat{X}_\lambda)$ will therefore be equal to 0 with probability $\mathbb{P}_{\mathbf{H}_0}(\lambda_0(Y) < \lambda)$. For $0 < \varepsilon < 1$, consider a threshold λ_ε that satisfies $\mathbb{P}_{\mathbf{H}_0}(\lambda_0(Y) < \lambda_\varepsilon) > 1 - \varepsilon$. We define the following test procedure of level ε , to which we refer as the thresholding test. We compute the estimator $\hat{X}_{\lambda_\varepsilon}$, and accept \mathbf{H}_0 if $\mathcal{T}(\hat{X}_{\lambda_\varepsilon}) = 0$, otherwise we reject it. This thresholding test is an alternative to the χ^2 test for independence. We compare the levels of the two tests in Section 5.3.

In practice we do not have access to the distribution $\mathbb{P}_{\mathbf{H}_0}(\lambda_0(Y) < \lambda)$, but we use the following procedure based on the parametric bootstrap (Efron, 1979) to compute a proxy for the threshold λ_ε , that we denote by λ_{QUT} . For a given observation matrix Y , estimate \hat{X}_0 , generate M_1 Poisson matrices $(Y_\ell)_{\ell=1}^{M_1} \sim \mathcal{P}(\exp(\hat{X}_0))$ and set $\lambda_{\text{QUT}}(Y)$ to the $(1 - \varepsilon)$ quantile of the distribution of $(m_1 m_2)^{-1} \left\| \mathcal{T}(Y_\ell - \exp(\hat{X}_0)) \right\|_{\sigma, \infty}$. Under the null hypothesis $\mathcal{T}(\bar{X}) = 0$, assume \hat{X}_0 is consistent, we obtain $\mathcal{T}(\hat{X}_{\lambda_{\text{QUT}}(Y)}) = 0$ with asymptotic probability $1 - \varepsilon$. In the experiments we see that $\lambda_{\text{QUT}}(Y)$ proves useful to select the rank of the interaction $\mathcal{T}(\bar{X})$.

5 Experiments

To assess the performance of our procedure we first consider synthetic data. We start by generating a contingency table according to the model $Y \sim \mathcal{P}(\exp(\bar{X}))$ with $\bar{X} = \bar{X}_0 + \bar{\Theta}$, $(\bar{X}_0)_{ij} = \bar{\alpha}_i + \bar{\beta}_j$. The row and column effects $\bar{\alpha}_i$ and $\bar{\beta}_j$ are drawn uniformly and we generate the interaction $\bar{\Theta}$ of rank K as $\bar{\Theta} = UDV^T$ with random orthonormal matrices $U = (u_{ij})$ and $V = (v_{ij})$, $D \in \mathbb{R}^{K \times K}$ being a diagonal matrix with the singular values of $\bar{\Theta}$ on its diagonal. The parameters of our simulation are the size of \bar{X} ($m_1 \times m_2$), the rank K of $\bar{\Theta}$ and the ratio of the nuclear norm of the interaction $\bar{\Theta}$ to the nuclear norm of the

additive part \bar{X}_0 : we define the signal to noise ratio (SNR) as $\|\bar{\Theta}\|_{\sigma,1} / \|\bar{X}_0\|_{\sigma,1}$.

5.1 Empirical assessment of CV and QUT

To compare the two methods for choosing λ described in Section 4, we consider a representative setting with $m_1 = 20$, $m_2 = 15$ and $K = 3$. Figure 1 represents the L_2 error of recovery between the estimated matrix \hat{X}_λ and the true parameter matrix \bar{X} as a function of λ . The maximum likelihood estimation in the independence model ($\bar{\Theta} = 0$) can be used as a benchmark. When λ is close to 0 we recover the saturated model while as λ increases we tend to the independence model. The rank of the estimator $\mathcal{T}(\hat{X}_\lambda)$ (number of singular values above $5 \cdot 10^{-6}$) decreases with λ . The two proposed procedures for choosing λ prove useful: λ_{QUT} selects the correct rank ($K = 3$) for the interaction and cross-validation achieves the best prediction error. An alternative procedure is a two-step approach where we fit the MLE with the rank found by QUT. Table 1 compares the estimated models based on the L_2 loss and the rank.

method	L_2 loss	rank
CV	7.37	10
QUT	13.38	3
MLE independence RC(0)	19.51	0
MLE oracle rank RC(3)	9.96	3
2 steps: MLE QUT rank RC(3)	9.96	3

Table 1: L_2 loss and rank for $m_1 = 20$, $m_2 = 15$ and $K = 3$.

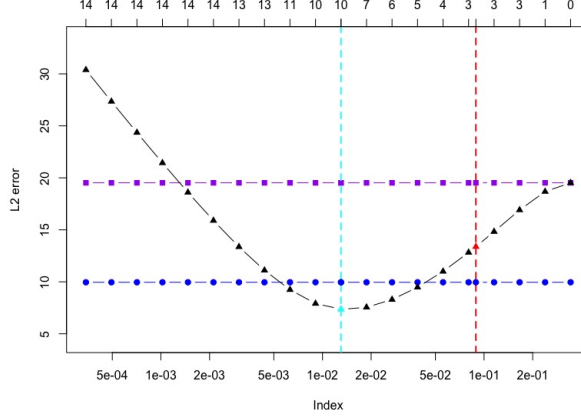


Figure 1: L_2 loss (black triangles) of ADMM estimator for $\lambda \in [1e - 4, 10]$ $m_1 = 20$, $m_2 = 15$, $K = 3$. Comparison of λ_{CV} (cyan dashed line) and λ_{QUT} (red dashed line) with the independence model (purple squares) and the MLE with oracle rank (blue points). The rank of the interaction is written on the top axis for every λ .

5.2 GAMMIT estimation

We compare our estimator in terms of L_2 error to the maximum likelihood estimators of the log-bilinear models with different ranks: the independence model $RC(0)$, the oracle rank $RC(K)$ and the rank estimated by QUT $RC(K_{QUT})$. We did not include cross-validation in our experiments because it was extremely costly in terms of computation time. For $K > 5$, the estimation of the $RC(K)$ model implemented in R ([Turner and Firth, 2015](#)) often fails, the errors of $RC(K)$ and $RC(K_{QUT})$ are therefore sometimes missing. Figure 2 highlights three interaction regimes. We start by looking at the rank 2 interaction (top three plots). In the small interaction regime (Figure 2 top left, $SNR = 0.2$), the interaction is too small to be distinguished from the Poisson noise, such that the independence model

achieves a better performance than $\text{RC}(K)$ and GAMMIT. The rank selected by QUT is of 1, and we see that the error of $\text{RC}(1)$ is very close to that of $\text{RC}(0)$. In the medium interaction regime (Figure 2 top center, $\text{SNR} = 0.7$) we recover the correct rank of 2 with

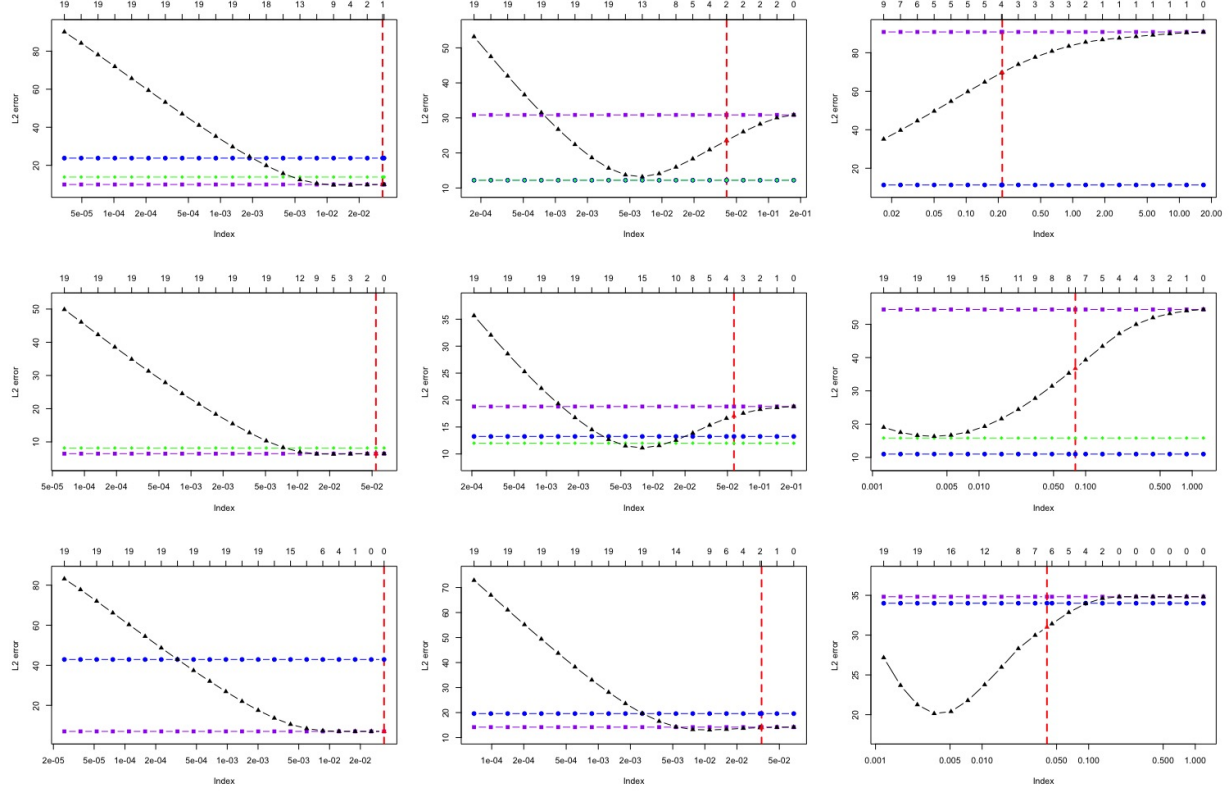


Figure 2: 50×20 matrices. Comparison of the L_2 error of GAMMIT (black triangles) with the independence model (purple squares), the rank oracle $\text{RC}(K)$ model (blue points) and the $\text{RC}(K_{\text{QUT}})$ (green diamonds). Results are drawn for a grid of λ with λ_{QUT} (red dot). The rank of the interaction is written on the top axis for every λ . Top $K = 2$, middle $K = 5$, bottom $K = 10$. From left to right $\text{SNR} = 0.2, 0.7, 1.7$.

QUT but obtain a highest error than the oracle RC(2). These two situations suggest to use GAMMIT with the two-step procedure. In the high interaction setting (Figure 2, top right SNR = 1.7), QUT overestimates the rank (here six instead of two), for which RC fails to calculate the maximum likelihood estimation (possibly because of numerical issues that occur in the available R libraries). In the medium rank setting ($K = 5$) we observe a very similar behavior. Looking at the high rank setting ($K = 10$), we observe similar results with an underestimation of the rank with QUT.

We further assess the rank recovery with a Monte-Carlo simulation. Table 2 gives the mode of the rank recovered with λ_{QUT} in various settings. These simulations provide good insights into the regimes for which λ_{QUT} is well suited: moderate interaction regimes with small ranks. The method tends to underestimate the rank when it is equal to 5. Unsurprisingly this tendency is exacerbated in the more difficult case when the rank is 10.

size	rank	SNR					
		0.2	0.3	0.4	0.5	0.7	1.7
50×20	2	0	2	2	2	2	4
50×20	5	1	2	2	4	4	4
50×20	10	0	0	1	1	2	5

Table 2: Mode of the estimated rank over 100 simulations using λ_{QUT} for different interaction intensities.

5.3 Thresholding test

We now perform the thresholding test defined in Section 4.2 and compare it to the χ^2 test for tables of size 50×20 . λ_{QUT} is computed as described in Section 4.2 with $\varepsilon = 0.05$ and generating $M_1 = 1e5$ Poisson matrices. The procedure is repeated $M_2 = 1e5$ times and the results are given in Table 3, for increasing values of the total number of counts N in the contingency table. As N increases we recover the asymptotic regime where the test has level 0.05. The results of the thresholding test are very similar to those of the χ^2 test. These first results highlight the potential of this approach in a testing setting and encourage further investigation.

N	chisq	thresh
13	1.00	1.00
673	0.95	0.96
4537	0.95	0.95
89556	0.95	0.94
990027	0.95	0.95

Table 3: Comparison of the levels of the thresholding and χ^2 tests for $M_1 = M_2 = 1e5$.

6 Data analyses

We start by demonstrating the interpretability of the GAMMIT interaction estimate on the *Death* data dataset and compare our results with that of the log-bilinear model. We then show how our method handles explanatory variables on an ecological example.

6.1 Causes of mortality per age

The contingency table *Death* (available at <http://factominer.free.fr/book/death.csv>) described in Husson et al. (2010) crosses causes of death with age categories for the French population in the year 2006. Age is encoded as a categorical variable with 12 categories (0 – 1, 1 – 4, 5 – 14, etc.) and 65 possible mortality causes are considered. A cell of the table therefore contains the number of people who died from a particular cause in a particular age category during the year 2006. GAMMIT is applied on this dataset, using row and column indicators as covariates (as in model (1.1)) and the threshold λ_{QUT} for the regularization parameter. We compare the results with the RC(3) analysis; the rank $K = 3$ was selected according to previous analyses on this data (Husson et al., 2010). We represent on Figure 3 biplot visualizations of the data in the two first dimensions of interaction. Models such as log-bilinear models provide a distance interpretation as follows. Two age categories or two mortality causes that are close to one another have similar profiles in the contingency table whereas an age category and a mortality cause that are close interact highly. More details about interpretation rules for these models can be found in Fithian and Josse (2017). Note also that correspondence analysis (Greenacre, 2007) can be applied to visualize such contingency tables. Biplot representation of correspondence analysis (not shown here) leads to interpretations which are very similar to those obtained with the RC model (Fithian and Josse, 2017). The interaction coefficients estimated with RC(3) are very large in amplitude and largest coefficients correspond to rare events. In particular, the age category 0 – 1 and related mortality causes such as *Sudden infant death syndrome* and *Complications in pregnancy and childbirth* completely drive the first dimension. Our regularized approach prevents such behaviors. The effect of the 0 – 1 category is also visible but shrunk, which reveals other important interactions that are not observed on the RC(3) biplot. The shape of the biplot and in particular the structure of the age categories (in red)

is known as the Guttman or horseshoe effect (Diaconis et al., 2008). The smallest distances found with GAMMIT concern the youngest age categories. In particular *Congenital defects* are very close to the children age categories, while *Road accidents* and *Alcohol abuse* are very close to the young categories 15 – 24 and 25 – 34.

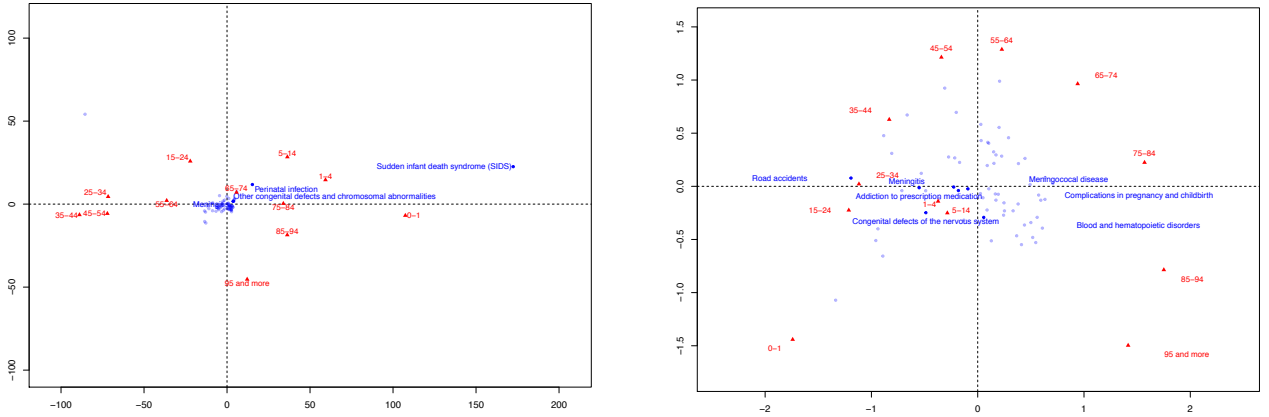


Figure 3: Visualization of the age categories (red) and mortality causes with largest contributions (blue) in the two first dimensions of interaction with the RC(3) model (left) and GAMMIT (right).

6.2 Distribution of Alpine plants in Aravo

Our second example is an ecological dataset that counts the abundance of 82 species of Alpine plants in 75 sites in France. Species traits providing physical information about the plants (height, spread, etc.) as well as environmental variables about the geography and climate of the different sites are also available. The data was initially published in Choler (2005) where ecologists looked for links between shifts in plant traits and environmental characteristics. The purpose of this analysis is to assess how incorporating covariates in our

model impacts the interpretation. We first apply GAMMIT without using the covariates, and obtain a rank of 3 for the interaction matrix.

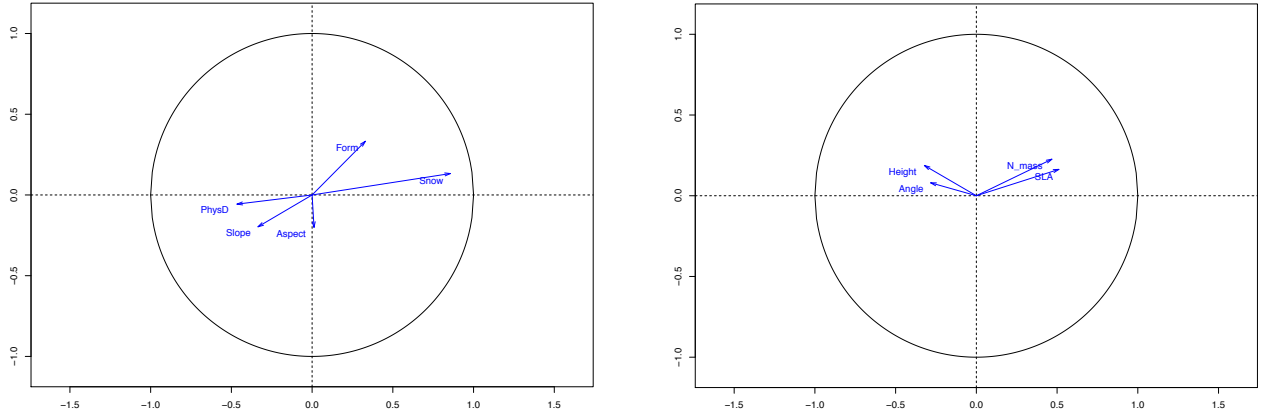


Figure 4: Correlation between environment (left) and species (right) covariates with the two first GAMMIT dimensions.

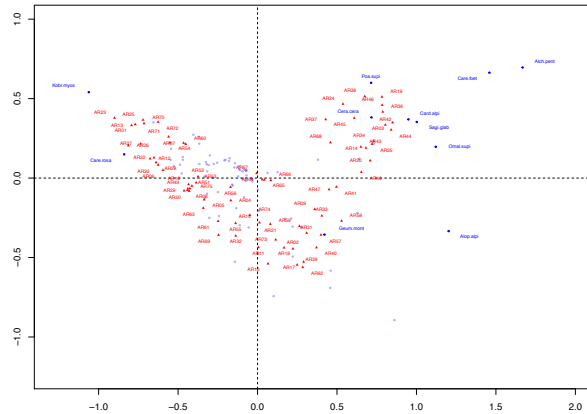


Figure 5: Scatterplot in the two first dimensions of interaction with GAMMIT with row-column indices as covariates of the environments (red) and 16 species (blue) with highest contribution to the dimensions.

Figure 4 (left) shows that environment covariates (not used in the analysis) and the two first directions of interaction are highly correlated: the first direction with the amount of *Snow* and the variable *PhysD* (which denotes the amount of physical disturbance causing unvegetated soil); the second with the *Aspect* variable (which denotes the compass direction, i.e. south, north, etc. that the site faced). On Figure 6 the first direction therefore separates environments with respect to the amount of snow, while the second direction separates the environments with respect to the compass direction. Similarly, the species covariates are highly correlated with the estimated directions of interaction (Figure 4, right): on Figure 6 the first direction separates the plants with respect to their *SLA* (Specific Leaf Area, defined as the ratio of the leaf surface to its dry mass) and to their mass based Nitrogen content (*Nmass*).

We then applied GAMMIT again, this time using the given traits as known covariates in our model. That is, we define R as the environment covariates and C as the species traits. The obtained results are very much interpretable and prove that we successfully separate the effect of the covariates from the interaction term. Indeed, the correlations between the known traits and the interaction directions is reduced by a factor between 3 and 10 (they are now too small to be represented on a plot). The rank of the estimated interaction matrix (using λ_{QUT}) is 1, which suggests that an additional variable, other than the measured explanatory covariates, summarizes the remaining interactions. Note that this is the first method available to select the rank in such models. Since the rank is 1 we cannot use a biplot representation, but we can compare the distances between species and environments before and after discarding the main effects. Figure 6 shows the species and environments that have the 10 highest interactions (smallest distances on the biplot), for the GAMMIT model without using the covariates (left) and the GAMMIT incorporating the covariates (right). We see that the species and environments involved differ, which

shows that our procedure could possibly lead to new interpretations. In particular, after incorporating the covariates, we extract species-environment couples much more clearly.

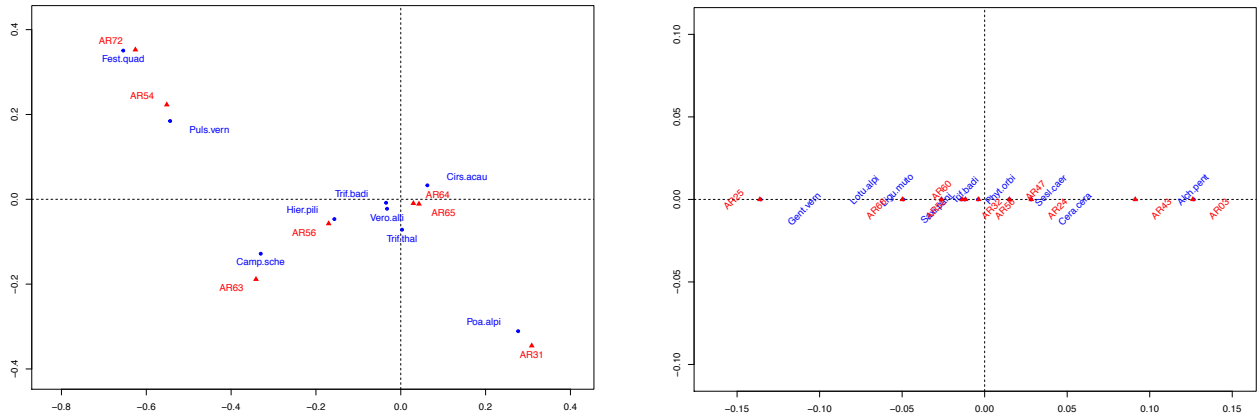


Figure 6: Visualization of the 10 largest interactions between environments (blue) and species (red) in the two first dimensions of interaction with GAMMIT for row-column indices (left) and explanatory covariates (right).

7 Discussion

We have proposed a regularized extension of the log-bilinear model for the analysis of contingency tables. We proved a theoretical upper bound for the estimation risk of our method and provided an algorithm that is available as an R code. We suggested two approaches to estimate the regularization parameter and showed how the results of our method can be interpreted through biplot representations. We finish by discussing some opportunities for further research.

First, the prediction performances of our method could possibly be improved by penalizing the singular values σ_i of Θ with different regularization parameters λ_i , such that

λ_i increases with i in the vein of adaptive lasso, as it was suggested in [Josse and Sardy \(2015b\)](#); [Gavish and Donoho \(2014b\)](#). Indeed, using a unique λ tends to shrink too much the first singular values which can lead to high errors. Second, since we do not regularize the main covariate effects, our method might have convergence issues when the contingency table has many zero counts. To prevent this, an option would be to add regularization in the \bar{X}_0 parameter.

There appears to be many potential extensions of GAMMIT. One possibility would be to use GAMMIT as a method to impute count data. We have indeed built an EM algorithm for cross-validation that handles and imputes missing values. GAMMIT could be a competitive alternative to single imputation methods for contingency tables. Directions of investigation are also to reduce the computational cost of this procedure for instance using approximations of the cross-validation and to extend the theoretical guarantees to the missing data setting. We are also eager to investigate further the thresholding test that we defined with the QUT procedure. In particular the power of the test should be assessed for different interaction settings. This could be a way to evaluate the difficulty of the problem of detecting and estimating interactions. The test statistic that we defined is not pivotal since it depends on the estimate of \bar{X}_0 under the independence model. This is an important issue, since the power of the test will depend on the quality of these estimates as was pointed out in [Giacobino et al. \(2016\)](#), and we wish to investigate the construction of a pivotal test statistic to overcome this issue. Finally, we would like to consider other sparsity inducing penalties. In particular, penalizing the Poisson log-likelihood by the absolute values of the coefficients of interaction matrix Θ could possibly lead to solutions where some interactions are driven to 0 and a small number of large interactions is selected. We are definitely interested in comparing GAMMIT with the results that would be obtained in this setting.

Acknowledgments

We thank Edgar Dobriban from the Statistics department of Stanford University for his careful reading and comments that helped improve this manuscript.

A Proof of Theorem 3.1

For sake of clarity, we write in the sequel \hat{X} instead of \hat{X}_λ . We begin the proof of Theorem 3.1 by some preparatory notations and lemmas. Given a matrix $X^1 \in \mathbb{R}^{m_1 \times m_2}$, we denote $\mathcal{S}_1(X^1)$ (resp. $\mathcal{S}_2(X^1)$) the span of left (resp. right) singular vectors of X^1 . We first define the orthogonal projection operator $\mathcal{P}_{X^1}^\perp : X^2 \mapsto P_{\mathcal{S}_1(X^1)}^\perp X^2 P_{\mathcal{S}_2(X^1)}^\perp$, where $P_{\mathcal{S}_1(X^1)}^\perp$ (resp. $P_{\mathcal{S}_2(X^1)}^\perp$) is the orthogonal projector on $\mathcal{S}_1(X^1)^\perp$ (resp. $\mathcal{S}_2(X^1)^\perp$). We also define $\mathcal{P}_{X^1} : X^2 \mapsto X^2 - P_{\mathcal{S}_1(X^1)}^\perp X^2 P_{\mathcal{S}_2(X^1)}^\perp$.

Lemma A.1. *For $X \in \mathbb{R}^{m_1 \times m_2}$*

- (i) $\left\| \mathcal{T}(\bar{X}) + \mathcal{P}_{\mathcal{T}(\bar{X})}^\perp(\mathcal{T}(\bar{X})) \right\|_{\sigma,1} = \left\| \mathcal{T}(\bar{X}) \right\|_{\sigma,1} + \left\| \mathcal{P}_{\mathcal{T}(\bar{X})}^\perp(\mathcal{T}(\bar{X})) \right\|_{\sigma,1},$
- (ii) $\left\| \mathcal{T}(\bar{X}) \right\|_{\sigma,1} - \left\| \mathcal{T}(X) \right\|_{\sigma,1} \leq \left\| \mathcal{P}_{\mathcal{T}(\bar{X})}(\mathcal{T}(X - \bar{X})) \right\|_{\sigma,1} - \left\| \mathcal{P}_{\mathcal{T}(\bar{X})}^\perp(\mathcal{T}(X - \bar{X})) \right\|_{\sigma,1},$
- (iii) $\left\| \mathcal{P}_{\mathcal{T}(\bar{X})}(\mathcal{T}(X - \bar{X})) \right\|_{\sigma,1} \leq \sqrt{2\text{rk}(\mathcal{T}(\bar{X}))} \left\| X - \bar{X} \right\|_{\sigma,2}.$

Proof. (i) The definition of $\mathcal{P}_{\mathcal{T}(\bar{X})}$ implies that the singular vector spaces of $\mathcal{T}(\bar{X})$ and of $\mathcal{P}_{\mathcal{T}(\bar{X})}^\perp(\mathcal{T}(\bar{X}))$ are orthogonal. Therefore

$$\left\| \mathcal{T}(\bar{X}) + \mathcal{P}_{\mathcal{T}(\bar{X})}^\perp(\mathcal{T}(\bar{X})) \right\|_{\sigma,1} = \left\| \mathcal{T}(\bar{X}) \right\|_{\sigma,1} + \left\| \mathcal{P}_{\mathcal{T}(\bar{X})}^\perp(\mathcal{T}(\bar{X})) \right\|_{\sigma,1}.$$

(ii) Writing $\mathcal{T}(X) = \mathcal{T}(\bar{X}) + \mathcal{P}_{\mathcal{T}(\bar{X})}^\perp(\mathcal{T}(X - \bar{X})) + \mathcal{P}_{\mathcal{T}(\bar{X})}(\mathcal{T}(X - \bar{X}))$ we get

$$\left\| \mathcal{T}(X) \right\|_{\sigma,1} \geq \left\| \mathcal{T}(\bar{X}) \right\|_{\sigma,1} + \left\| \mathcal{P}_{\mathcal{T}(\bar{X})}^\perp(\mathcal{T}(X - \bar{X})) \right\|_{\sigma,1} - \left\| \mathcal{P}_{\mathcal{T}(\bar{X})}(\mathcal{T}(X - \bar{X})) \right\|_{\sigma,1},$$

using the triangular inequality and the orthonormality of the left and right singular vector spaces of $\mathcal{T}(\bar{X})$ and $\mathcal{P}_{\mathcal{T}(\bar{X})}^\perp(\mathcal{T}(X - \bar{X}))$. This shows that

$$\|\mathcal{T}(\bar{X})\|_{\sigma,1} - \|\mathcal{T}(X)\|_{\sigma,1} \leq \|\mathcal{P}_{\mathcal{T}(\bar{X})}(\mathcal{T}(X - \bar{X}))\|_{\sigma,1} - \left\| \mathcal{P}_{\mathcal{T}(\bar{X})}^\perp(\mathcal{T}(X - \bar{X})) \right\|_{\sigma,1}. \quad (\text{A.1})$$

(iii) For all $X \in \mathbb{R}^{m_1 \times m_2}$, $\mathcal{P}_{\mathcal{T}(\bar{X})}(\mathcal{T}(X)) = P_{S_1(\mathcal{T}(\bar{X}))} \mathcal{T}(X - \bar{X}) P_{S_2(\mathcal{T}(\bar{X}))}^\perp + \mathcal{T}(X - \bar{X}) P_{S_2(\mathcal{T}(\bar{X}))}$ implies that $\text{rk}(\mathcal{P}_{\mathcal{T}(\bar{X})}(\mathcal{T}(X - \bar{X}))) \leq 2\text{rk}(\mathcal{T}(\bar{X}))$. This and the Cauchy-Schwarz inequality give

$$\begin{aligned} \|\mathcal{P}_{\mathcal{T}(\bar{X})}(\mathcal{T}(X - \bar{X}))\|_{\sigma,1} &\leq \sqrt{2\text{rk}(\mathcal{T}(\bar{X}))} \|\mathcal{T}(X - \bar{X})\|_{\sigma,2} \\ &\leq \sqrt{2\text{rk}(\mathcal{T}(\bar{X}))} \|X - \bar{X}\|_{\sigma,2}, \end{aligned}$$

□

Lemma A.2. Assume $\lambda > 2 \|\nabla \Phi_Y(\bar{X})\|_{\sigma,\infty}$. Then,

$$(i) \left\| P_{\mathcal{T}(\bar{X})}^\perp \mathcal{T}(\bar{X} - \hat{X}) \right\|_{\sigma,1} \leq 3 \left\| P_{\mathcal{T}(\bar{X})} \mathcal{T}(\bar{X} - \hat{X}) \right\|_{\sigma,1} + \left\| (I - \mathcal{T})(\bar{X} - \hat{X}) \right\|_{\sigma,1},$$

$$(ii) \left\| \hat{X} - \bar{X} \right\|_{\sigma,1} \leq \left(1 + \sqrt{32\text{rk}(\mathcal{T}(\bar{X}))} + \sqrt{2} \right) \left\| \bar{X} - \hat{X} \right\|_{\sigma,2}.$$

Proof. (i) The convexity of Φ_Y and the duality of the norms $\|\cdot\|_{\sigma,\infty}$ and $\|\cdot\|_{\sigma,1}$ (Boyd and Vandenberghe, 2004, Section 2.6) yield

$$\begin{aligned} \Phi_Y(\hat{X}) - \Phi_Y(\bar{X}) &\geq \langle \nabla \Phi_Y(\bar{X}), \hat{X} - \bar{X} \rangle \\ &\geq -\|\nabla \Phi_Y(\bar{X})\|_{\sigma,\infty} \left\| \hat{X} - \bar{X} \right\|_{\sigma,1} \geq -\frac{\lambda}{2} \left\| \hat{X} - \bar{X} \right\|_{\sigma,1}. \end{aligned}$$

Since by definition of \hat{X} , $\lambda \left(\left\| \mathcal{T}(\hat{X}) \right\|_{\sigma,1} - \left\| \mathcal{T}(\bar{X}) \right\|_{\sigma,1} \right) \leq -\Phi_Y(\hat{X}) + \Phi_Y(\bar{X})$, we get

$$\begin{aligned} \lambda \left(\left\| \mathcal{T}(\hat{X}) \right\|_{\sigma,1} - \left\| \mathcal{T}(\bar{X}) \right\|_{\sigma,1} \right) &\leq \frac{\lambda}{2} \left\| \hat{X} - \bar{X} \right\|_{\sigma,1} \\ &\leq \frac{\lambda}{2} \left(\left\| \mathcal{T}(\hat{X} - \bar{X}) \right\|_{\sigma,1} + \left\| (I - \mathcal{T})(\hat{X} - \bar{X}) \right\|_{\sigma,1} \right). \end{aligned}$$

Reusing [A.1](#) this gives

$$\begin{aligned} & \lambda \left(\left\| \mathcal{P}_{\mathcal{T}(\bar{X})}^\perp \mathcal{T}(\hat{X} - \bar{X}) \right\|_{\sigma,1} - \left\| \mathcal{P}_{\mathcal{T}(\bar{X})} \mathcal{T}(\hat{X} - \bar{X}) \right\|_{\sigma,1} \right) \\ & \leq \frac{\lambda}{2} \left(\left\| \mathcal{P}_{\mathcal{T}(\bar{X})}^\perp \mathcal{T}(\hat{X} - \bar{X}) \right\|_{\sigma,1} + \left\| \mathcal{P}_{\mathcal{T}(\bar{X})} \mathcal{T}(\hat{X} - \bar{X}) \right\|_{\sigma,1} + \left\| (I - \mathcal{T})(\hat{X} - \bar{X}) \right\|_{\sigma,1} \right), \end{aligned}$$

and finally

$$\left\| \mathcal{P}_{\mathcal{T}(\bar{X})}^\perp \mathcal{T}(\hat{X} - \bar{X}) \right\|_{\sigma,1} \leq 3 \left\| \mathcal{P}_{\mathcal{T}(\bar{X})} \mathcal{T}(\hat{X} - \bar{X}) \right\|_{\sigma,1} + \left\| (I - \mathcal{T})(\hat{X} - \bar{X}) \right\|_{\sigma,1},$$

(ii) Now, Lemma [A.1](#) (iii) and the triangular inequality give

$$\begin{aligned} \left\| \mathcal{T}(\hat{X} - \bar{X}) \right\|_{\sigma,1} & \leq 4 \left\| \mathcal{P}_{\mathcal{T}(\bar{X})} \mathcal{T}(\hat{X} - \bar{X}) \right\|_{\sigma,1} + \left\| (I - \mathcal{T})(\hat{X} - \bar{X}) \right\|_{\sigma,1} \\ & \leq \left(\sqrt{32\text{rk}(\mathcal{T}(\bar{X}))} + \sqrt{2} \right) \left\| \bar{X} - \hat{X} \right\|_{\sigma,2}. \end{aligned} \tag{A.2}$$

We now use Cauchy's interlacing theorem ([Bhatia \(1997\)](#) Corollary III.1.5) which states that, denoting $\sigma_k(X)$, $1 \leq k \leq K$ the k^{th} singular value of X in the decreasing order, P_1, P_2 two orthogonal projections and $Z = P_2 X P_1$, then $\sigma_1(X) \geq \sigma_1(Z) \geq \sigma_2(X) \geq \dots \geq \sigma_{K-1}(Z) \geq \sigma_K(X)$. Since $\mathcal{T}(X) = \Pi_2 X \Pi_1$, with Π_1 and Π_2 orthogonal projectors, this implies that

$$\left\| \hat{X} - \bar{X} \right\|_{\sigma,1} \leq \left\| \mathcal{T}(\hat{X} - \bar{X}) \right\|_{\sigma,1} + \left\| \hat{X} - \bar{X} \right\|_{\sigma,\infty} \leq \left\| \mathcal{T}(\hat{X} - \bar{X}) \right\|_{\sigma,1} + \left\| \hat{X} - \bar{X} \right\|_{\sigma,2}.$$

Combining this with [\(A.2\)](#) we obtain

$$\left\| \hat{X} - \bar{X} \right\|_{\sigma,1} \leq \left(1 + \sqrt{32\text{rk}(\mathcal{T}(\bar{X}))} + \sqrt{2} \right) \left\| \bar{X} - \hat{X} \right\|_{\sigma,2},$$

which proves (ii). □

We now proceed to the proof of Theorem 3.1. The proof derives from two main arguments using the strong convexity of Φ_Y and the empirical Bregman divergence

$$D_{\Phi_Y}(\hat{X}, \bar{X}) = \Phi_Y(\hat{X}) - \Phi_Y(\bar{X}) - \langle \nabla \Phi_Y(\bar{X}), \hat{X} - \bar{X} \rangle. \quad (\text{A.3})$$

On the one hand, the strong convexity of Φ_Y implies, $\mu \left\| \bar{X} - \hat{X} \right\|_{\sigma,2}^2 / (2m_1 m_2) \leq D_{\Phi_Y}(\hat{X}, \bar{X})$. On the other hand, by definition of the estimator \hat{X} we have

$$\Phi_Y(\hat{X}) - \Phi_Y(\bar{X}) \leq \lambda \left(\left\| \mathcal{T}(\bar{X}) \right\|_{\sigma,1} - \left\| \mathcal{T}(\hat{X}) \right\|_{\sigma,1} \right).$$

Subtracting $\langle \nabla \Phi_Y(\bar{X}), \hat{X} - \bar{X} \rangle$ on both sides we get

$$\mu \frac{\left\| \bar{X} - \hat{X} \right\|_{\sigma,2}^2}{2m_1 m_2} \leq -\langle \nabla \Phi_Y(\bar{X}), \hat{X} - \bar{X} \rangle + \lambda \left(\left\| \mathcal{T}(\bar{X}) \right\|_{\sigma,1} - \left\| \mathcal{T}(\hat{X}) \right\|_{\sigma,1} \right).$$

We now upper bound the two terms in the right hand side of the previous relation. The duality of $\|\cdot\|_{\sigma,1}$ and $\|\cdot\|_{\sigma,\infty}$ yields $-\langle \nabla \Phi_Y(\bar{X}), \bar{X} - \hat{X} \rangle \leq \left\| \nabla \Phi_Y(\bar{X}) \right\|_{\sigma,\infty} \left\| \hat{X} - \bar{X} \right\|_{\sigma,1}$.

$$-\langle \nabla \Phi_Y(\bar{X}), \hat{X} - \bar{X} \rangle \leq \left\| \nabla \Phi_Y(\bar{X}) \right\|_{\sigma,\infty} \times \left(\left\| \mathcal{P}_{\mathcal{T}(\bar{X})}(\mathcal{T}(\hat{X} - \bar{X})) \right\|_{\sigma,1} + \left\| \mathcal{P}_{\mathcal{T}(\bar{X})}^\perp(\mathcal{T}(\hat{X} - \bar{X})) \right\|_{\sigma,1} + \left\| (I - \mathcal{T})(\bar{X} - \hat{X}) \right\|_{\sigma,1} \right), \quad (\text{A.4})$$

which gives an upper bound for the first term. To bound the second term, we apply Lemma A.1(ii) to \hat{X} , which gives

$$\left\| \mathcal{T}(\bar{X}) \right\|_{\sigma,1} - \left\| \mathcal{T}(\hat{X}) \right\|_{\sigma,1} \leq \left\| \mathcal{P}_{\mathcal{T}(\bar{X})}(\mathcal{T}(\hat{X} - \bar{X})) \right\|_{\sigma,1} - \left\| \mathcal{P}_{\mathcal{T}(\bar{X})}^\perp(\mathcal{T}(\hat{X} - \bar{X})) \right\|_{\sigma,1}. \quad (\text{A.5})$$

Combining Equation (A.4) and Equation (A.5) we obtain

$$\begin{aligned} \mu \frac{\left\| \bar{X} - \hat{X} \right\|_{\sigma,2}^2}{2m_1 m_2} &\leq \left(\left\| \nabla \Phi_Y(\bar{X}) \right\|_{\sigma,\infty} + \lambda \right) \left\| \mathcal{P}_{\mathcal{T}(\bar{X})}(\mathcal{T}(\hat{X} - \bar{X})) \right\|_{\sigma,1} \\ &\quad + \left(\left\| \nabla \Phi_Y(\bar{X}) \right\|_{\sigma,\infty} - \lambda \right) \left\| \mathcal{P}_{\mathcal{T}(\bar{X})}^\perp(\mathcal{T}(\hat{X} - \bar{X})) \right\|_{\sigma,1} \\ &\quad + \left\| \nabla \Phi_Y(\bar{X}) \right\|_{\sigma,\infty} \left\| (I - \mathcal{T})(\bar{X} - \hat{X}) \right\|_{\sigma,1}, \end{aligned} \quad (\text{A.6})$$

and $\lambda \geq 2 \left\| \nabla \Phi_Y(\bar{X}) \right\|_{\sigma, \infty}$ ensures

$$\mu \frac{\left\| \bar{X} - \hat{X} \right\|^2}{m_1 m_2} \leq 3\lambda \left\| \mathcal{P}_{\mathcal{T}(\bar{X})}(\mathcal{T}(\hat{X} - \bar{X})) \right\|_{\sigma, 1} + \lambda \left\| (I - \mathcal{T})(\bar{X} - \hat{X}) \right\|_{\sigma, 1}. \quad (\text{A.7})$$

Since $\text{rk}((I - \mathcal{T})(\bar{X} - \hat{X})) \leq K_1 + K_2$ and $\left\| (I - \mathcal{T})(\bar{X} - \hat{X}) \right\|_{\sigma, 2} \leq \left\| \bar{X} - \hat{X} \right\|_{\sigma, 2}$ we have $\left\| (I - \mathcal{T})(\bar{X} - \hat{X}) \right\|_{\sigma, 1} \leq \sqrt{K_1 + K_2} \left\| \bar{X} - \hat{X} \right\|_{\sigma, 2}$, which together with Item (iii) and using $2(a^2 + b^2) \geq (a + b)^2$ yields Theorem 3.1.

B Proof of Theorem 3.2

Consider the random matrices defined by $Z_{ij} = (y_{ij} - \exp(\bar{x}_{ij}))E_{ij}$, $i, j \in [m_1] \times [m_2]$, with E_{ij} is the (i, j) th canonical matrix. With this notation $\nabla \Phi_Y = (m_1 m_2)^{-1} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} Z_{ij}$, $\mathbb{E}[Z_{ij}] = 0$, $\underline{\sigma} \leq \mathbb{E}[\left\| Z_{ij} Z_{ij}^T \right\|_{\sigma, \infty}] \leq \bar{\sigma}$ and $\underline{\sigma} \leq \mathbb{E}[\left\| Z_{ij}^T Z_{ij} \right\|_{\sigma, \infty}] \leq \bar{\sigma}$ for all $i, j \in [m_1] \times [m_2]$. We define the quantity

$$\sigma_Z^2 = \max \left(\frac{1}{m_1 m_2} \left\| \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}[Z_{ij} Z_{ij}^T] \right\|_{\sigma, \infty}, \frac{1}{m_1 m_2} \left\| \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mathbb{E}[Z_{ij}^T Z_{ij}] \right\|_{\sigma, \infty} \right). \quad (\text{B.1})$$

Using **H2** there exists $K > 0$ such that for all i, j , $\mathbb{E}[\exp(\left\| Z_{ij} \right\|_{\sigma, \infty} / K)] < +\infty$. We apply (Klopp, 2014, Proposition 11). There exists a constant $c_K < \infty$ that depends only on K such that for all $t > 0$, with probability at least $1 - e^{-t}$ we have

$$\frac{1}{m_1 m_2} \left\| \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} Z_{ij} \right\|_{\sigma, \infty} \leq c_K \max \left\{ \sigma_Z \sqrt{\frac{t + \log(d)}{m_1 m_2}}, K \left(\log \frac{K}{\sigma_Z} \right) \frac{t + \log(d)}{m_1 m_2} \right\}, \quad (\text{B.2})$$

where σ_Z is defined as in (B.1). Since $M/(m_1 m_2) \underline{\sigma}^2 \leq \sigma_Z^2 \leq M/(m_1 m_2) \bar{\sigma}^2$ (where $\underline{\sigma}$ and $\bar{\sigma}$ are defined in **H1**) and setting $t = \log(d)$, we obtain that with probability at least $1 - d^{-1}$

$$\left\| \nabla \Phi_Y(\bar{X}) \right\|_{\sigma, \infty} \leq c_K \max \left\{ \bar{\sigma} \frac{\sqrt{2M \log(d)}}{m_1 m_2}, K \left(\log \sqrt{m} \frac{K}{\underline{\sigma}} \right) \frac{2 \log(d)}{m_1 m_2} \right\}, \quad (\text{B.3})$$

with $M \geq d/2$ and $m \leq d/2$. Asymptotically in d the left term dominates, and assuming $d \geq d^* = \max \{K^2(2\bar{\sigma}\bar{\sigma}^2)^{-1}, (4K^2/\bar{\sigma}^2)^4\}$ ensures that $\sqrt{d \log(d)}\bar{\sigma} \geq K \left(\log \sqrt{d/2} K/\bar{\sigma} \right) 2 \log(d)$. Taking $\lambda = 2c\bar{\sigma} \sqrt{2M \log(d)}/(m_1 m_2)$ then guarantees $\lambda \geq \|\nabla \Phi_Y\|_{\sigma, \infty}$ with probability at least $1 - d^{-1}$ which completes the proof.

C Proof of Proposition 1

We write $X = X_0 + \Theta$ with $\Theta = \mathcal{T}(X)$. The zero thresholding statistics is given by the minimum value of λ , $\lambda_0(Y)$, that sets $\hat{\Theta}_\lambda$ to 0 under the null hypothesis:

$$\lambda_0(Y) = \min_{\lambda} 0 \in \partial \{ \Phi_Y^\lambda(X_0, \Theta) + 1_{\mathcal{K}}(X_0 + \Theta) \} |_{\Theta=0},$$

where $1_{\mathcal{K}}(X_0 + \Theta)$ is the indicator of \mathcal{K} equal to 0 on \mathcal{K} and $+\infty$ elsewhere. On the one hand, under the constraint $\Theta = 0$ we get $\hat{X}_0 = \underset{X \in \mathcal{K}, \mathcal{T}(X)=0}{\operatorname{argmin}} \Phi_Y(X)$. On the other hand, the subdifferential of the objective function Φ_Y^λ with respect to Θ at $\Theta = 0$ is given by

$$\partial_{\Theta} \Phi_Y^\lambda |_{\Theta=0} = -\frac{1}{m_1 m_2} (Y - \exp(X_0)) + \lambda \partial_{\Theta} \|\Theta\|_{\sigma, 1} |_{\Theta=0} + \partial_{\Theta} 1_{\mathcal{K}}(X_0 + \Theta) |_{\Theta=0}.$$

First Lemma C.1 guarantees that $0 \in \partial_{\Theta} 1_{\mathcal{K}}(X_0 + \Theta) |_{\Theta=0}$. Then Lemma C.2 ensures that $0 \in \partial \Phi_Y^\lambda(\Theta) |_{\Theta=0}$ (and therefore $\Theta = 0$ is solution of the optimization problem) if and only if

$$0 \in -\frac{1}{m_1 m_2} (Y - \exp(\hat{X}_0)) + \lambda W,$$

$\|\mathcal{T}(W)\|_{\sigma, \infty} < 1$. Which is itself equivalent to $(I - \mathcal{T})(W) = (I - \mathcal{T}) \left(-(m_1 m_2)^{-1} (Y - \exp(\hat{X}_0)) \right)$ and $\lambda \geq (m_1 m_2)^{-1} \left\| \mathcal{T}(Y - \exp(\hat{X}_0)) \right\|_{\sigma, \infty}$.

Lemma C.1. *Let \mathcal{C} be a compact set, $\mathcal{C}_{\mathcal{T}}$ be the image of \mathcal{C} by projector \mathcal{T} , and $1_{\mathcal{C}}$ be the indicator of \mathcal{C} equal to 0 on \mathcal{C} and $+\infty$ elsewhere. Consider a matrix $X \in \mathcal{C}$. Define $f : \mathcal{C}_{\mathcal{T}} \rightarrow \mathbb{R}_+$ be the function such that $f(A) = 1_{\mathcal{C}}(X + A)$. Then $0 \in \partial f(A) |_{A=0}$.*

Proof. The subdifferential of f is defined by

$$\partial f(A) = \{W \in \mathbb{R}^{m_1 \times m_2}, f(B) \geq f(A) + \langle W, (B - A) \rangle; B \in \mathcal{C}_T\}.$$

We now fix $A = 0$. $X \in \mathcal{C}$ therefore $f(0) = 0$. Let B be a matrix of \mathcal{C}_T . If $X + B \in \mathcal{C}$ then $f(B) = 0$ and $f(B) \geq f(0) + \langle 0, (B - A) \rangle$. If $X + B \notin \mathcal{C}$ then $f(B) = +\infty$ and $f(B) \geq f(0) + \langle 0, (B - A) \rangle$. Therefore $0 \in \partial f(A)|_{A=0}$. \square

Lemma C.2. *Let $g : \mathcal{V}^\perp \rightarrow \mathbb{R}_+$ be the function defined by $g(A) = \|A\|_{\sigma,1}$ for $A \in \mathcal{V}^\perp$. Then $\partial g(0) = \left\{ W \in \mathbb{R}^{m_1 \times m_2}, \|\mathcal{T}(W)\|_{\sigma,\infty} < 1 \right\}$.*

Proof. Recall the definition of the subdifferential

$$\partial g(A) = \{W \in \mathbb{R}^{m_1 \times m_2}, g(B) \geq g(A) + \langle W, (B - A) \rangle; B \in \mathcal{V}^\perp\}.$$

We now fix $A = 0$. Let us decompose $W = W_1 + W_2$, $W_1 \in \mathcal{V}$ and $W_2 = \mathcal{T}(W) \in \mathcal{V}^\perp$. Since $\langle W, B \rangle = \langle W_2, B \rangle$, $\|W_2\|_{\sigma,\infty} \leq 1$ is a sufficient condition for $W \in \partial g(0)$. Now assume $\|W_2\|_{\sigma,\infty} > 1$. Let $W_2 = U\Sigma V^T$, where $\Sigma_{11} > 1$ is the largest singular value of W_2 , U and V are orthogonal matrices of left and right singular vectors. Let us define $B = U\tilde{\Sigma}V^T$, $\tilde{\Sigma}_{11} = 1$ and $\tilde{\Sigma}_{ij} = 0$ elsewhere. Since $\mathcal{T}(B) = (I - \Pi_1)B(I - \Pi_2)$, Lemma C.3 ensures $B \in \mathcal{V}^\perp$. We have $g(B) = 1$ and $\langle W_2, B \rangle = \Sigma_{11} > g(B)$. Therefore $\|W_2\|_{\sigma,\infty} > 1 \Rightarrow W \notin \partial g(0)$, from which we conclude

$$\partial g(0) = \left\{ W \in \mathbb{R}^{m_1 \times m_2}, \|\mathcal{T}(W)\|_{\sigma,\infty} < 1 \right\}.$$

\square

Lemma C.3. *Let P_1 and P_2 be two orthogonal projectors on subspaces of \mathbb{R}^{m_1} and \mathbb{R}^{m_2} , respectively. Let $A \in \mathbb{R}^{m_1 \times m_2}$ be a matrix such that $A = P_1 A P_2$ and let $B \in \mathbb{R}^{m_1 \times m_2}$ be a matrix such that $\text{Im}(B) \subseteq \text{Im}(A)$ and $\text{Im}(B^T) \subseteq \text{Im}(A^T)$. Then $B = P_1 B P_2$.*

Proof. $A = P_1 A P_2$ implies that $\text{Im}(A) \subseteq \text{Im}(P_1)^\perp$ and $\text{Im}(A^T) \subseteq \text{Im}(P_2)^\perp$. Therefore $\text{Im}(B) \subseteq \text{Im}(P_1)^\perp$ and $\text{Im}(B^T) \subseteq \text{Im}(P_2)^\perp$ and $B = P_1 B P_2$. \square

References

- Agresti, A. (2013). *Categorical Data Analysis, 3rd Edition*. Wiley.
- Bhatia, R. (1997). *Matrix Analysis*. Springer.
- Bhatia, R. and F. Kittaneh (2000). Cartesian decompositions and Schatten norms. *Linear Algebra and its Applications* 318(1), 109 – 116.
- Bigot, J., C. Deledalle, and D. FÃral (2016). Generalized sure for optimal shrinkage of singular values in low-rank matrix denoising. *arXiv:1605.07412*.
- Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* 3(1), 1–22.
- Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. New York, NY, USA: Cambridge University Press.
- Cai, J.-F., E. J. Candès, and Z. Shen (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* 20(4), 1956–1982.
- Cao, Y. and Y. Xie (2016, March). Poisson matrix recovery and completion. *IEEE Transactions on Signal Processing* 64(6).
- Chambolle, A. and T. Pock (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision* 40(1), 120–145.
- Choler, P. (2005, 1). Consistent shifts in alpine plant traits along a mesotopographical gradient. *Arctic, Antarctic, and Alpine Research* 37(4), 444–453.

- Christensen, R. (2010). *Log-Linear Models*. Springer-Verlag, New York.
- Collins, M., S. Dasgupta, and R. Schapire (2001). A generalization of principal component analysis to the exponential family. In *Advances in Neural Information Processing Systems*. MIT Press.
- Davenport, M. A., Y. Plan, E. van den Berg, and M. Wootters (2012, September). 1-Bit Matrix Completion. *ArXiv e-prints*.
- de Falguerolles, A. (1998). Log-bilinear biplot in action. In J. Blasius and M. . Greenacre (Eds.), *Visualisation of categorical data*, pp. 527–533. Academic Press.
- de Leeuw, J. (2006). Principal component analysis of binary data by iterated singular value decomposition. *Computational Statistics and Data Analysis* 50(1), 21–39.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society B* 39(1), 1–38.
- Diaconis, P., S. Goel, and S. Holmes (2008). Horseshoes in multidimensional scaling and local kernel methods. *Annals of Applied Statistics* 2(3), 777–807.
- Donoho, D. L. and I. M. Johnstone (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika* 81, 425–455.
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics*, 1–26.
- Feuerverger, A., Y. He, and S. Khatri (2012, 05). Statistical significance of the netflix challenge. *Statist. Sci.* 27(2), 202–231.

- Figueiredo, M. A. T. and J. M. Bioucas-Dias (2010, December). Restoration of poissonian images using alternating direction optimization. *Trans. Img. Proc.* 19(12), 3133–3145.
- Fithian, W. and J. Josse (2017). Multiple correspondence analysis & the multilogit bilinear model. *Journal of Multivariate Analysis*.
- Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani (2007, 12). Pathwise coordinate optimization. *Ann. Appl. Stat.* 1(2), 302–332.
- Gavish, M. and D. L. Donoho (2014a). The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory* 60(8).
- Gavish, M. and D. L. Donoho (2014b). Optimal shrinkage of singular values. *arXiv:1405.7511*.
- Giacobino, C., S. Sardy, J. Diaz Rodriguez, and N. Hengardner (2016). Quantile universal threshold for model selection. *arXiv:1511.05433v2*.
- Glowinski, R. and A. Marrocco (1974). Sur l’approximation, par éléments finis d’ordre 1, et la résolution, par pénalisation-dualité, d’une classe de problèmes de Dirichlet non linéaires. *C. R. Acad. Sci. Paris Sér. A* 278, 1649–1652.
- Goodman, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Annals of Statistics* 13, 10–69.
- Gower, J., S. Lubbe, and N. le Roux (2011). *Understanding Biplots*. John Wiley & Sons.
- Greenacre, M. J. (2007). *Correspondence Analysis in Practice, Second Edition*. Chapman & Hall.

- Hastie, T., R. Mazumder, J. Lee, and R. Zadeh (2014, October). Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. *ArXiv e-prints*.
- Hastie, T., R. Mazumder, J. Lee, and R. Zadeh (2015). Matrix completion and low-rank svd via fast alternating least squares. *Journal in Machine Learning Research*.
- Husson, F., S. Lê, and J. Pagès (2010). *Exploratory Multivariate Analysis by Example Using R*. Chapman & Hall/CRC.
- Jeong, T., H. Woo, and S. Yun (2013). Frame-based poisson image restoration using a proximal linearized alternating direction method. *Inverse Problems* 29(7), 075007.
- Josse, J. and S. Sardy (2015a). Adaptive shrinkage of singular values. *Statistics and Computing*, 1–10.
- Josse, J. and S. Sardy (2015b). Adaptive shrinkage of singular values. *Statistics and Computing*, 1–10.
- Josse, J. and S. Wager (2016). Bootstrap-based regularization for low-rank matrix estimation. *Journal of Machine Learning Research* 17(124), 1–29.
- Kateri, M. (2014). *Contingency Table Analysis*. Springer New York.
- Klopp, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli* 20(1), 282–303.
- Li, J. and D. Tao (2013). Simple exponential family PCA. *IEEE Transactions on Neural Networks and Learning Systems* 24(3), 485–497.
- Little, R. J. A. and D. B. Rubin (1987, 2002). *Statistical Analysis with Missing Data*. New-York: John Wiley & Sons series in probability and statistics.

- Liu, L., E. Dobriban, and A. Singer (2016). epca: High dimensional exponential family pca. *arXiv:1611.05550*.
- Luisier, F., T. Blu, and M. Unser (2011). Image denoising in mixed poisson-gaussian noise. *IEEE Transactions on Image Processing* 20(3), 696–708.
- O. Stegle, S. A. T. and J. C. Marioni (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* 16(3), 133–145.
- Peres-Neto, P. R., S. Dray, and C. J. F. t. Braak (2016). Linking trait variation to the environment: critical issues with community-weighted mean correlation resolved by the fourth-corner approach. *Ecography*, n/a–n/a.
- Pierson, E. and Y. Christopher (2015). Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology* 16(1), 241.
- R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Raginsky, M., R. M. Willett, Z. T. Harmany, and R. F. Marcia (2010, Aug). Compressed sensing performance bounds under poisson noise. *IEEE Transactions on Signal Processing* 58(8), 3990–4002.
- Salmon, J., Z. Harmany, C. Deledalle, and R. Willett (2014). Poisson noise reduction with non-local pca. *Journal of Mathematical Imaging and Vision* 48(2), 279–294.
- Shabalin, A. A. and A. B. Nobel (2013). Reconstruction of a low-rank matrix in the presence of Gaussian noise. *Journal of Multivariate Analysis* 118, 67–76.

Turner, H. and D. Firth (2015). *Generalized nonlinear models in R: An overview of the gnm package*. R package version 1.0-8.